DR. ALVIN'S PUBLICATIONS

LINEAR REGRESSION USING EXCEL

BY DR. ALVIN ANG

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

CONTENTS

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

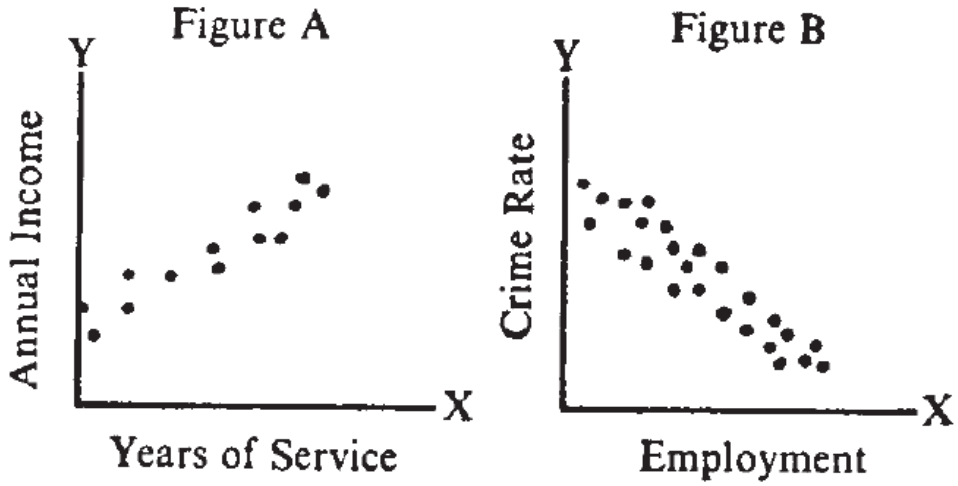
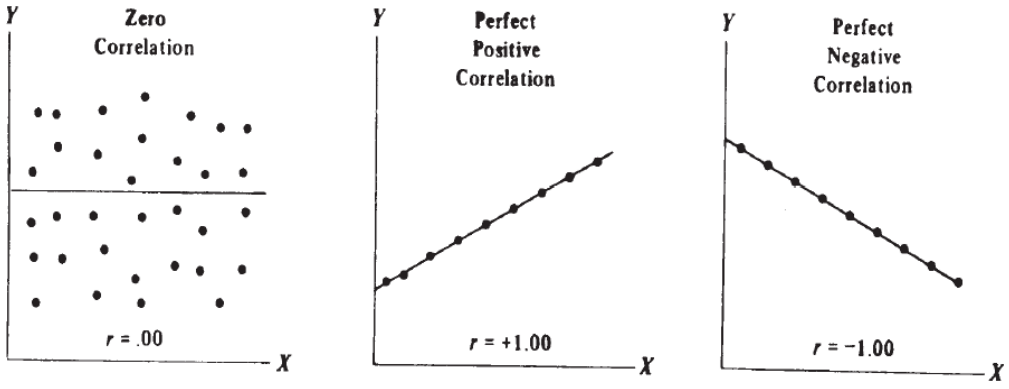


Figure 1: Scatter Plots which could be Linearly Regressed (SUSS 2016)

- Figure 1(A) shows a scatter plot that could be positively linearly regressed.
- Figure 1(B) shows a scatter plot that could be negatively linearly regressed.

The following scatter diagrams depict correlations of 0, +1.0, and -1.0.



COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

II. STEP 1: SCATTER PLOT TO CHECK LINEARITY

Table 1: Repair Cost ($) vs Age (years)

Repair Cost ($) - Y	Age (years) - X
170	1
130	1
180	2
205	2
220	3
243	3
290	4
275	4
404	5
380	5

- Does the data in Table 1 show a linear relationship?
- We perform a scatter plot in Excel to find out.

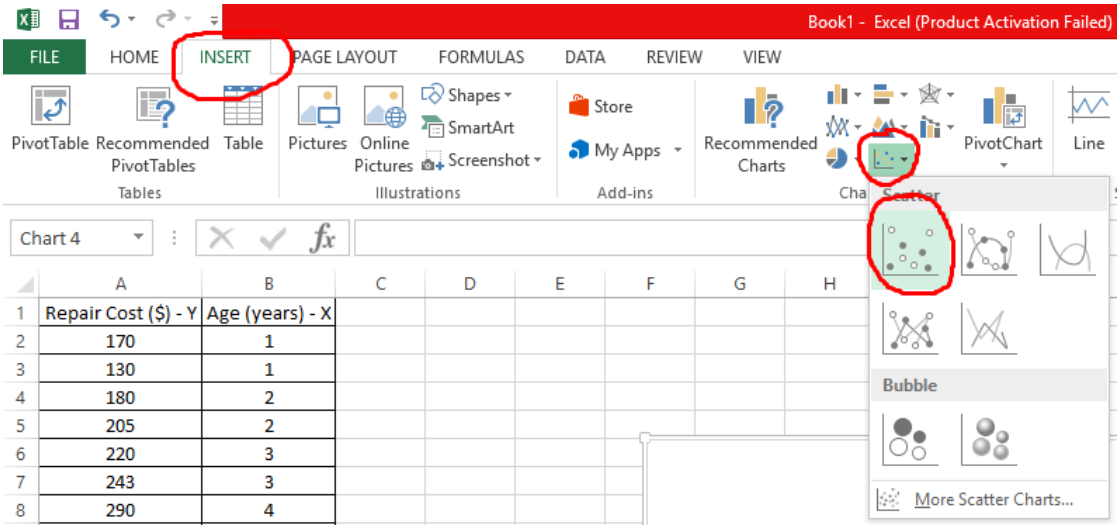


Figure 2: Scatter Plot

- Click Insert → Choose the 1st Scatter Plot

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

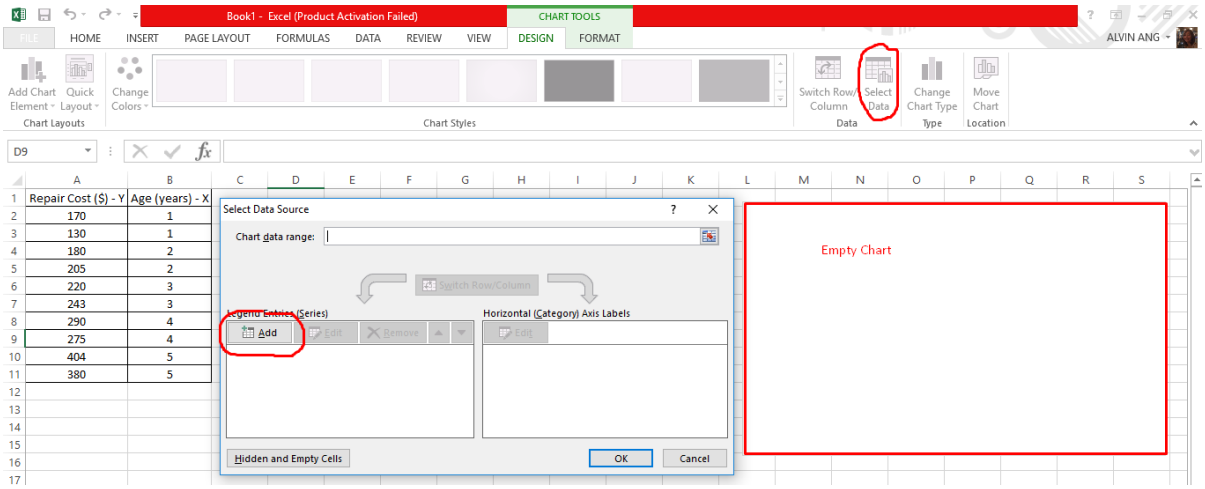
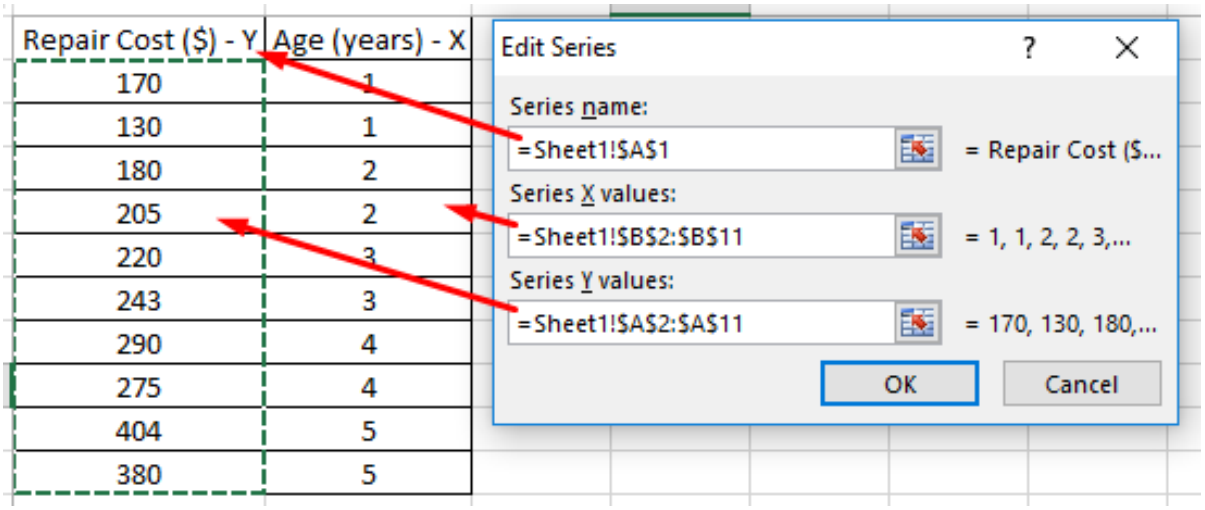


Figure 3: Label the Correct Axis

- Select the Empty Chart that popped up
- At the Design Tab, click "Select Data"
- Click "Add"



- Series Name: Select "Repair Cost" header
- X Values: Select Age (years) column
- Y Values: Select Repair Cost ($) column
- Click OK

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

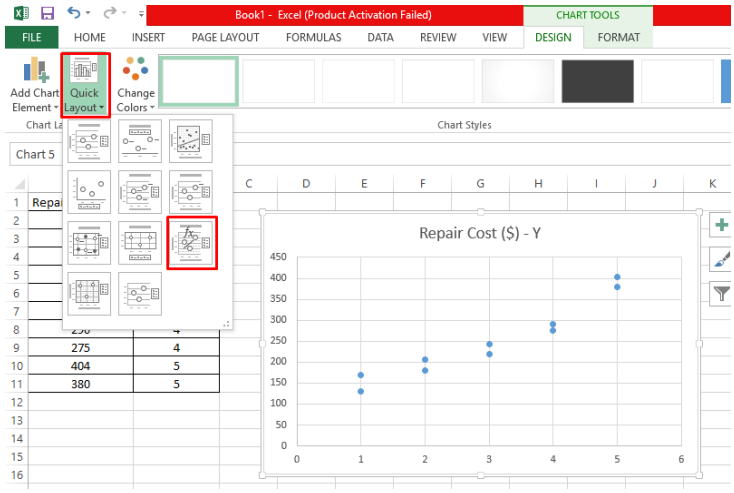


Figure 4: Editing using Quick Layout

- Click Quick Layout and select the fx graph.

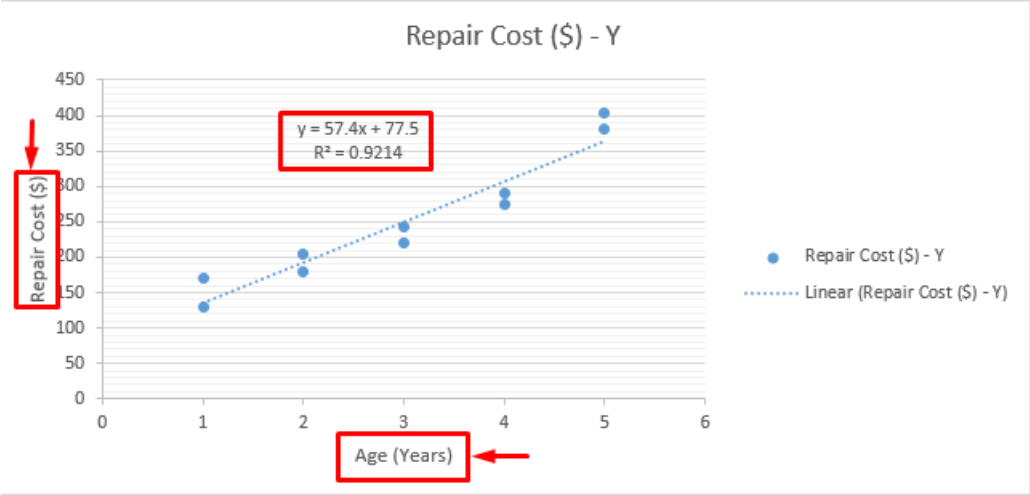
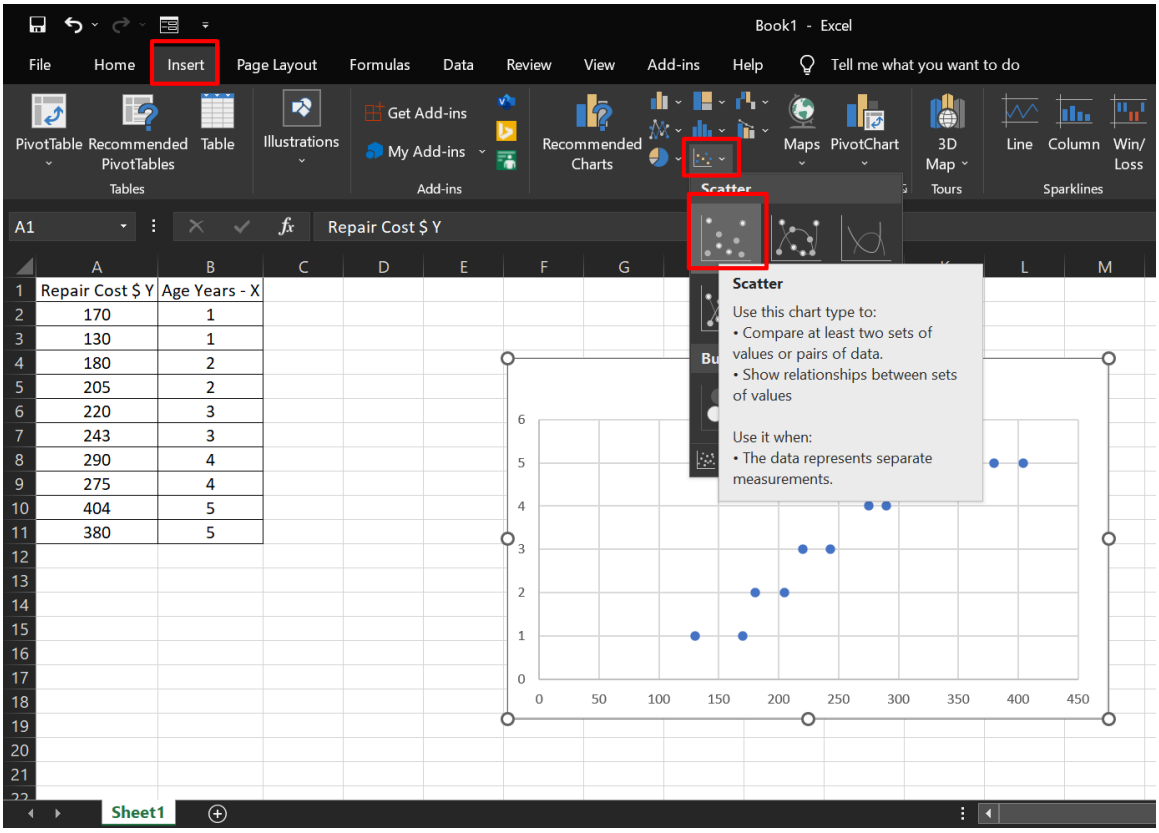


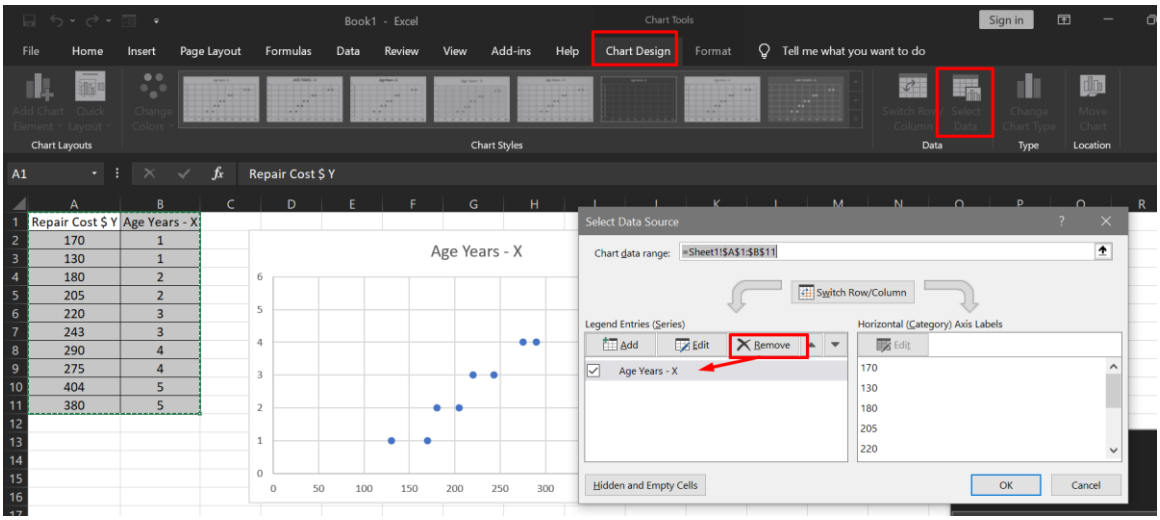
Figure 5: Final Scatter Plot with Linear Regression Line

- Figure 5 shows the final scatter plot with the linear regression line. This shows that the linearity check is OK.
- Remember to edit the Axis Titles.
- Note the Best Fit Line Equation given is $\hat{Y} = 57.4X + 77.5$

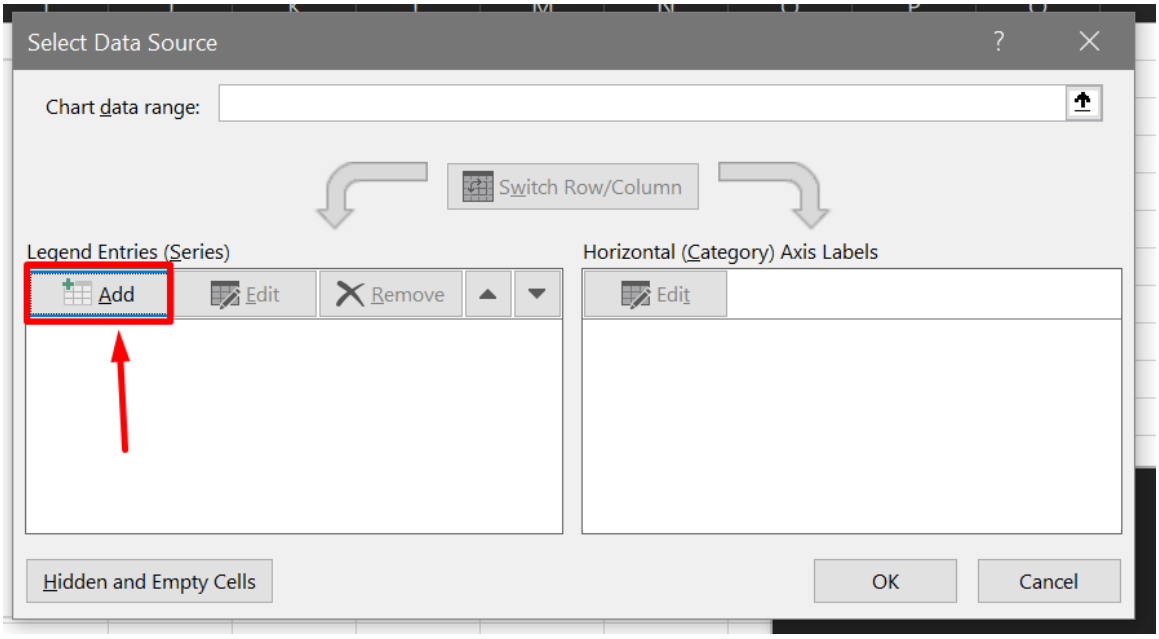
COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

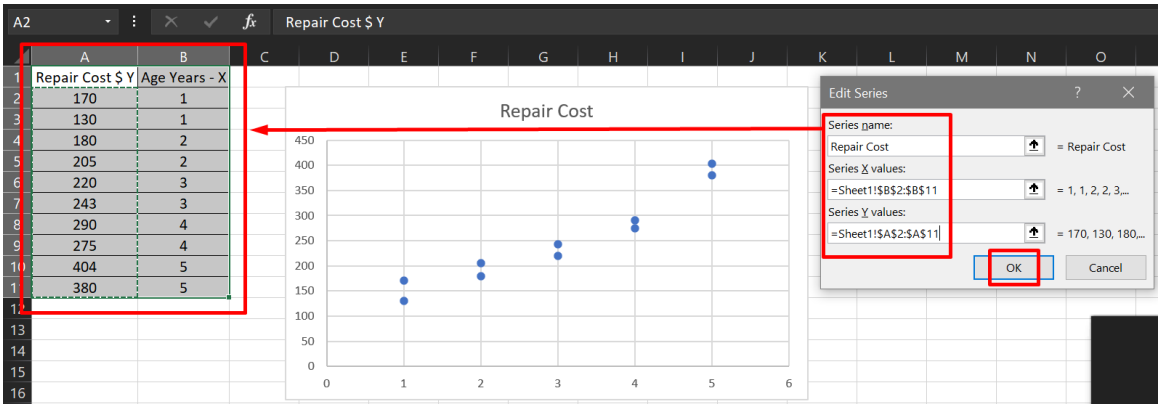
A. ALTERNATIVE WAY FOR CURVE FITTING

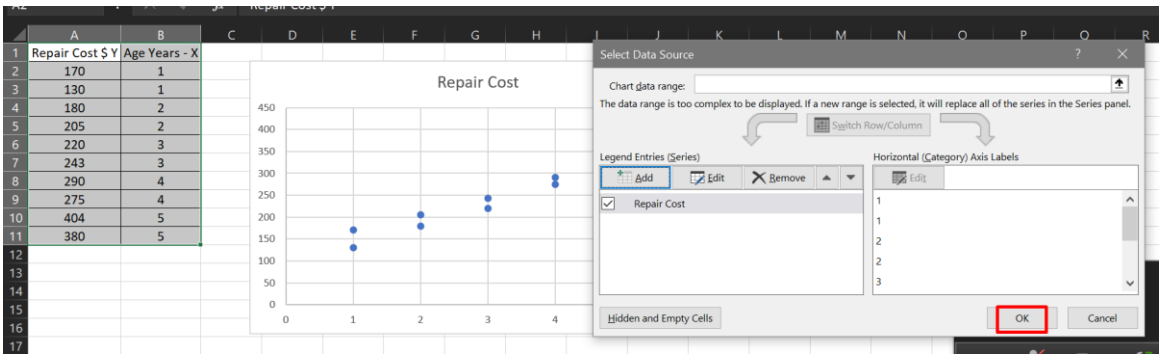




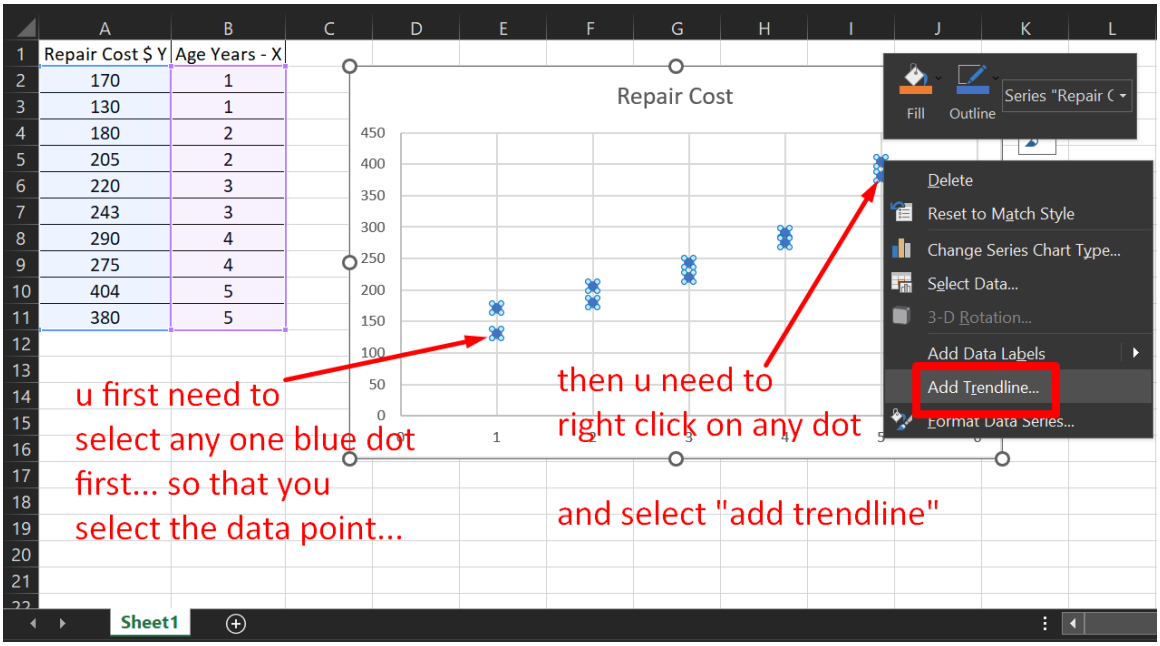
COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

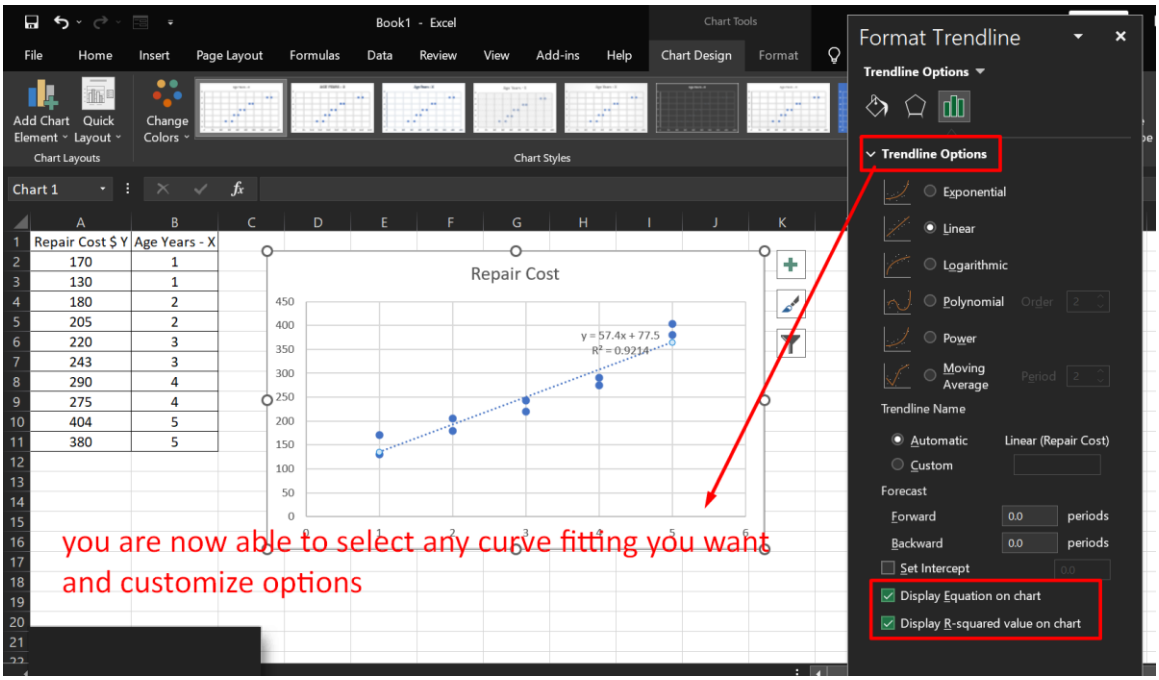






COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG





COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

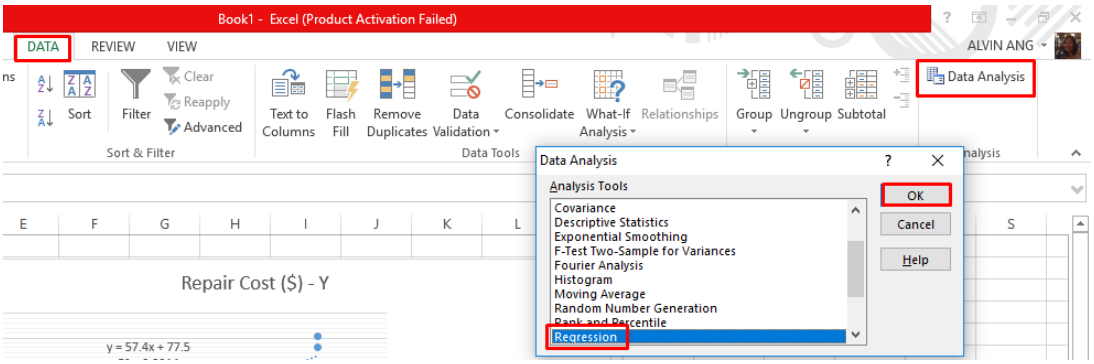


Figure 6: Click on Data Analysis

- In order to perform Regression Analysis, you first need to install Excel Analysis Toolpak.
- Please refer to Ang (2018) on how to install.
- Click the DATA tab → Data Analysis → Regression → OK.

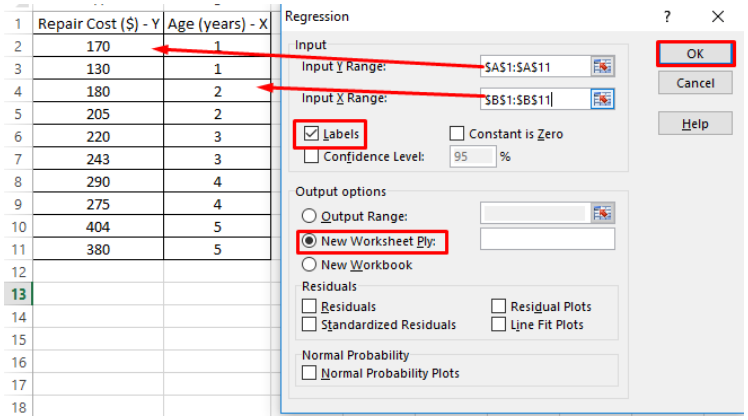


Figure 7: Regression Analysis Inputs

- Input Y Range: Repair Cost column
- Input X Range: Age column
- Select Labels
- Select New Worksheet Ply
- Click OK

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 8: Regression Analysis Output

- Figure 8 shows a new sheet created with all the Regression Analysis Output.
- We will explain it in detail in the next section.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

IV. STEP 3: ANALYZING THE REGRESSION ANALYSIS OUTPUT

A. MULTIPLE R

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 9: Multiple R

- Multiple R is also called : ***Correlation Coefficient*** & is also labelled as : r
- Multiple R or r shows the correlation between actual values of the dependent variable, Y and the predicted values for Y.
- Multiple R = r = 0 indicates no correlation
- Multiple R = r = 1 means perfect correlation.
- Multiple R = r = 0.960 suggests a strong positive correlation.
- This means that as age of the car increases, so does annual repair cost.
- Figure 10 shows the equation of Multiple R or r.
- Since we rarely use calculation by hand, we shall ignore the equation. (Figure 11)

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

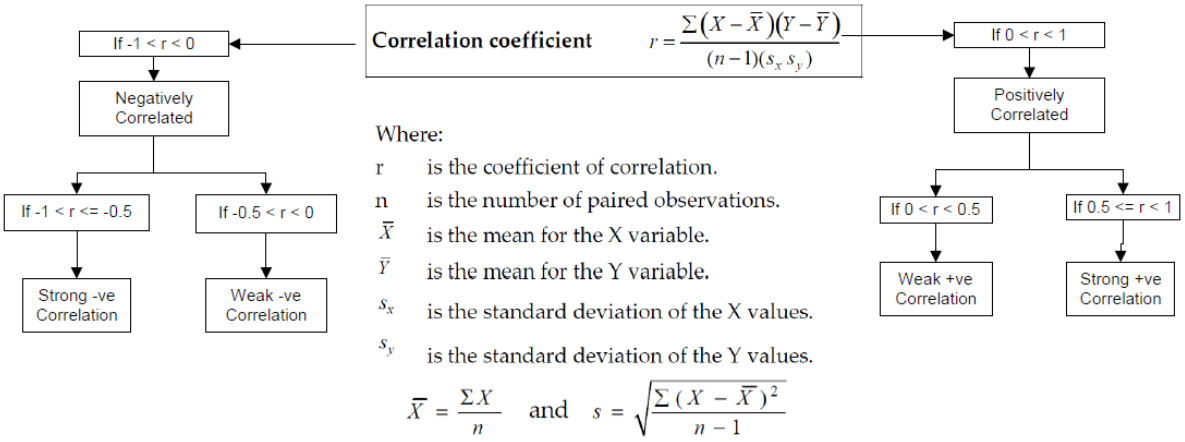


Figure 10: Multiple R = r = Correlation Coefficient

	A	B	C	D	E	F	G
	Repair Cost			Age			
	Y	$(Y-\bar{Y})$	$(Y-\bar{Y})^2$	X	$(X-\bar{X})$	$(X-\bar{X})^2$	$(X-\bar{X})(Y-\bar{Y})$
	$170	-80	6400	1	-2	4	160
	130	-120	14400	1	-2	4	240
	180	-70	4900	2	-1	1	70
	205	-45	2025	2	-1	1	45
	220	-30	900	3	0	0	0
	243	-7	49	3	0	0	0
	290	40	1600	4	1	1	40
	275	25	625	4	1	1	25
	404	154	23716	5	2	4	308
	380	130	16900	5	2	4	260
n	10			10			
Σ	2500	0.00	71515	30	0.00	20	1148

Step 1. Compute the means using sums in Column A and D:

$$\bar{Y} = \frac{\Sigma Y}{n} = \frac{2500}{10} = 250 \qquad \bar{X} = \frac{\Sigma X}{n} = \frac{30}{10} = 3.0$$

Step 2. Compute the standard deviations using the sums in Column C and F:

$$s_y = \sqrt{\frac{\Sigma(Y-\bar{Y})^2}{n-1}} = \sqrt{\frac{71515}{10-1}} = 89.14 \qquad s_x = \sqrt{\frac{\Sigma(X-\bar{X})^2}{n-1}} = \sqrt{\frac{20}{10-1}} = 1.49$$

Step 3. Compute the coefficient of correlation *r* using the formula, the sum from Column G in the table, and the calculated standard deviations:

$$r = \frac{\Sigma(X-\bar{X})(Y-\bar{Y})}{(n-1)(s_x s_y)} = \frac{1148}{9(1.4907)(89.14)} = \frac{1148}{1196.71} = 0.9599 = 0.960$$

Figure 11: How to Calculate Multiple R = r = by hand = not important

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

B. R SQUARE

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 12: R Square

- R Square is known as the ***Coefficient of Determination***
- In the case of Multiple Regression, it's called the Coefficient of ***Multiple*** Determination
- R^2 = (Multiple R) ^{2} = r^2 = (0.9599) ^{2} = 0.9214
- R^2 = the proportion of variation of Y accounted for by variation in X.
- For example, if R^2 = 0.92, that means that X (age) accounts for 92% of the variation of Y (repair cost).
- Since $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$, we shall describe what is SSR / SSE / SST in the next section.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

C. SSR / SSE / SST

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df		*SS*	*MS*	*F*	*Significance F*		
Regression	1	SSR	65895.2	SSR/165895.2	93.81936	1.07652E-05		
Residual	n-2 8	SSE	5618.9	702.3625	SSE/(n-2)			
Total	n-1 9	SST	71514.1					
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 13: SSR / SSE / SST ...

- SS: Sum of Squares
- SSR: Sum of Squares of Regression (variation) = 65895.2 = $\sum\left(\hat{Y}-\bar{Y}\right)^2$
  - Where $\hat{Y}=57.4X+77.5$ (this equation was given by Excel Scatter Plot when we used the fx graph in Figure 5).
  - Where $\bar{Y}=\frac{\sum Y}{n}=250$ (refer to Figure 11)
  - We obtain $\hat{Y}\ \&\ \bar{Y}$ through Table 2 below.
- SSE: Sum of Squares of Error (variation) = 5618.9 = $\sum\left(Y-\hat{Y}\right)^2$
- SST: Sum of Squares Total (variation) = 71514.1 = = SSR + SSE = $\sum\left(Y-\bar{Y}\right)^2$

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

- MSR: Mean of Squares of Regression = 702.3625 = $\frac{SSR}{1}$

- MSE: Mean of Squares of Error = 65895.2 = $\frac{SSE}{n-2}$

Table 2: Table to Obtain SSE / SSR / SST

$\hat{Y} = 57.4X + 77.5$

X	bX	$\hat{Y}$	Y	$(Y-\hat{Y})$	$(Y-\hat{Y})^2$
1	57.40	134.9	$170	35.1	1232.01
1	57.40	134.9	130	-4.9	24.01
2	114.80	192.3	180	-12.3	151.29
2	114.80	192.3	205	12.7	161.29
3	172.20	249.7	220	-29.7	882.09
3	172.20	249.7	243	-6.7	44.89
4	229.60	307.1	290	-17.1	292.41
4	229.60	307.1	278	-32.1	1030.41
5	287.00	364.5	404	39.5	1560.25
5	287.00	364.5	380	15.5	240.25
				$SUM =$	5618.9 SSE

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

D. DEGREES OF FREEDOM (DF)

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.959911184					
R Square	0.92142948					
Adjusted R Square	0.911608165					
Standard Error	26.50212256					
Observations	10					
ANOVA						
	df		*SS*	*MS*	*F*	*Significance F*
Regression		1	65895.2	65895.2	93.81936	1.07652E-05
Residual	n-2	8	5618.9	702.3625		
Total	n-1	9	71514.1			
	Coefficients		*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	77.5		19.65450012	3.943117328	0.004277	32.17664145	122.8233586
age (years)	57.4		5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678

Figure 14: Degrees of Freedom (df)

- The Degree of Freedom (df) for Regression is 1
    - Reason: $\hat{Y} = 57.4X + 77.5$
    - Y hat is dependent only on one X.
    - In simple words, there is only one way to get Y based on the Regression Line.
    - It means that I can predict every particular observation (Y: the regression line) based only on X
    - That's why only 1 df.
- The df for Residual is 8
    - This means that there are 8/10 ways to get a guess of the residual.
    - The degrees of freedom associated with the error term (the residual) is (n − 2)(SUSS 2016).

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

E. ADJUSTED R^2

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 15: Adjusted R Square

$$R^2_{adj} = 1 - \left[\frac{(1 - R^2)(n-1)}{n-k-1} \right]$$

Where:

- R^2_{adj} is needed because R^2 is not very accurate.
- R^2_{adj} is more effective than R2.
- Because as the number of independent variables,X_1, X_2, X_3 etc... increases, R^2 increases.
- This makes R^2 inaccurate.
- R^2_{adj} will not necessarily increase when a new variable is added.
- R^2_{adj} takes into account the effect of potential overfitting due to the number of independent variables.
- R^2: ***Coefficient of Determination***

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

- n: number of observations = 10
- k: number of X's (or number of independent variables).
- In this case, $\hat{Y} = 57.4X + 77.5$, which means that there is only one X (age) (k = 1).
- In the case of Multiple Regression where there are many Xs, k will increase.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

F. STANDARD ERROR

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 16: Standard Error of Estimate

$$S_{y.x} = \sqrt{\frac{SSE}{n-2}} = \sqrt{\frac{\sum\left(Y-\hat{Y}\right)^2}{n-2}}$$

- $S_{y.x}$: Standard Error of Estimate = 26.5
- $S_{y.x}$ is a measure of dispersion of values around the regression line
- $S_{y.x}$: is the Standard Deviation of the Residuals away from the proposed line.
- Figure 17 shows what Residuals are.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

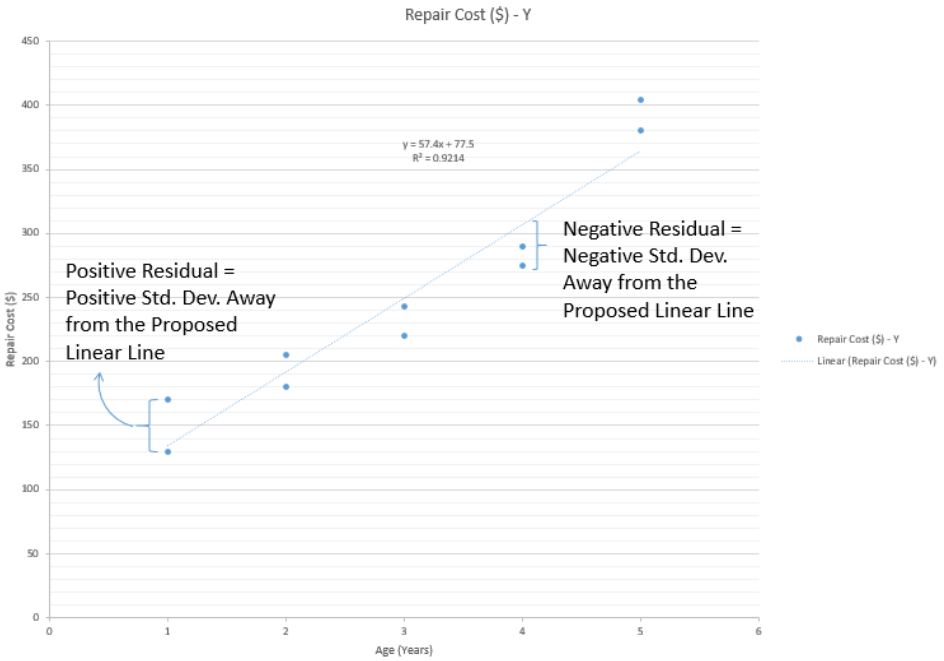
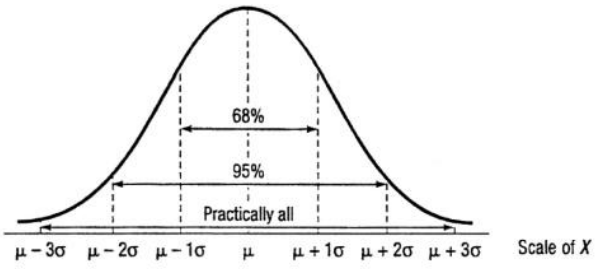


Figure 17: What are Residuals?

- Small $S_{y.x}$ = Lesser Scatter = Good predictor
- Big $S_{y.x}$ = More Scatter = Bad predictor
- Similar to Multiple R or r, both measures strength of relationship between X and Y
- But $S_{y.x}$ has same units as Y, Multiple R or r has range -1 to 1
- Since $S_{y.x}$ = 26.5, this shows that about 68% of the predictions should be within ±$26.50 (±1σ) of the actual repair costs and about 95% should be within ($26.50 x 2) = ±$53 (±2σ) of actual repair costs.



COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

G. COEFFICIENTS

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 18: Coefficients

- Recall earlier that $\hat{Y} = 57.4X + 77.5$
- The Y intercept (Repair Cost) = $$77.50.
- The gradient of X = 57.4.
- This shows that every increase of X (age) by 1 year
- → Y will increase linearly by Y = 57.4 (1) + 77.5 = $134.90.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

H. F STAT

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 19: F Stat vs t stat

- $F = \frac{MSR}{MSE} = \frac{65895.2}{702.3625} = 93.819$
- The purpose of F and Significance F is for Global Testing (to test for All X).
- In other words, the question (for multiple regression) is "Are ALL X important in this model?"
- Since this is a case of single variable linear regression, the question becomes "Is X (age) important in this model? Does Y (repair cost) really depend on it?"
- Let's create a Global Hypothesis test:
    - Null Hypothesis: $H_0 : \beta_1 = 0$ (X (age) is not important)
    - Alternate Hypothesis: $H_1 : \beta_1 \neq 0$ (X(age) is important)
    - Where β_1 is actually referring to X.
    - $\alpha = 5\%$

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

- Since df (Regression) = Numerator = 1
- Since df (Residual) = Denominator = 8
- Referring to Table 3, F critical (F_{crit}) = 5.32
- Referring to Figure 19, F statistic (F_{stat}) = 93.82
- Since $F_{crit} < F_{stat}$ → Accept H1 → X is important.

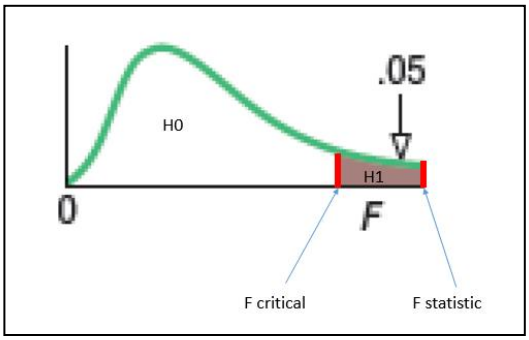
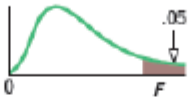


Table 3: F distribution for alpha = 5%

B.4 Critical Values of the *F* Distribution at a 5 Percent Level of Significance

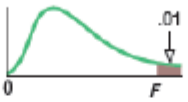


Degrees of Freedom for the Denominator	Degrees of Freedom for the Numerator															
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	161	200	216	225	230	234	237	239	241	242	244	246	248	249	250	251
2	18.5	19.0	19.2	19.2	19.3	19.3	19.4	19.4	19.4	19.4	19.4	19.4	19.4	19.5	19.5	19.5
3	10.1	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24	2.16	2.09	2.01	1.96	1.92	1.87
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16	2.09	2.01	1.93	1.89	1.84	1.79
40	4.08	3.23	2.84	2.61	2.45	2.34	2.25	2.18	2.12	2.08	2.00	1.92	1.84	1.79	1.74	1.69
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99	1.92	1.84	1.75	1.70	1.65	1.59
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91	1.83	1.75	1.66	1.61	1.55	1.50
∞	3.84	3.00	2.60	2.37	2.21	2.10	2.01	1.94	1.88	1.83	1.75	1.67	1.57	1.52	1.46	1.39

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

Table 4: F distribution for alpha = 1%

B.4 Critical Values of the *F* Distribution at a 1 Percent Level of Significance (*concluded*)



Degrees of Freedom for the Denominator	Degrees of Freedom for the Numerator															
	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40
1	4052	5000	5403	5625	5764	5859	5929	5981	6022	6056	6106	6157	6209	6235	6261	6287
2	98.5	99.0	99.2	99.2	99.3	99.3	99.4	99.4	99.4	99.4	99.4	99.4	99.4	99.5	99.5	99.5
3	34.1	30.8	29.5	28.7	28.2	27.9	27.7	27.5	27.3	27.2	27.1	26.9	26.7	26.6	26.5	26.4
4	21.2	18.0	16.7	16.0	15.5	15.2	15.0	14.8	14.7	14.5	14.4	14.2	14.0	13.9	13.8	13.7
5	16.3	13.3	12.1	11.4	11.0	10.7	10.5	10.3	10.2	10.1	9.89	9.72	9.55	9.47	9.38	9.29
6	13.7	10.9	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87	7.72	7.56	7.40	7.31	7.23	7.14
7	12.2	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62	6.47	6.31	6.16	6.07	5.99	5.91
8	11.3	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81	5.67	5.52	5.36	5.28	5.20	5.12
9	10.6	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26	5.11	4.96	4.81	4.73	4.65	4.57
10	10.0	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85	4.71	4.56	4.41	4.33	4.25	4.17
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54	4.40	4.25	4.10	4.02	3.94	3.86
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30	4.16	4.01	3.86	3.78	3.70	3.62
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10	3.96	3.82	3.66	3.59	3.51	3.43
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94	3.80	3.66	3.51	3.43	3.35	3.27
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80	3.67	3.52	3.37	3.29	3.21	3.13
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69	3.55	3.41	3.26	3.18	3.10	3.02
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59	3.46	3.31	3.16	3.08	3.00	2.92
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51	3.37	3.23	3.08	3.00	2.92	2.84
19	8.18	5.93	5.01	4.50	4.17	3.94	3.77	3.63	3.52	3.43	3.30	3.15	3.00	2.92	2.84	2.76
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37	3.23	3.09	2.94	2.86	2.78	2.69
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31	3.17	3.03	2.88	2.80	2.72	2.64
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26	3.12	2.98	2.83	2.75	2.67	2.58
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21	3.07	2.93	2.78	2.70	2.62	2.54
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17	3.03	2.89	2.74	2.66	2.58	2.49
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13	2.99	2.85	2.70	2.62	2.54	2.45
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98	2.84	2.70	2.55	2.47	2.39	2.30
40	7.31	5.18	4.31	3.83	3.51	3.29	3.12	2.99	2.89	2.80	2.66	2.52	2.37	2.29	2.20	2.11
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63	2.50	2.35	2.20	2.12	2.03	1.94
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47	2.34	2.19	2.03	1.95	1.86	1.76
∞	6.63	4.61	3.78	3.32	3.02	2.80	2.64	2.51	2.41	2.32	2.18	2.04	1.88	1.79	1.70	1.59

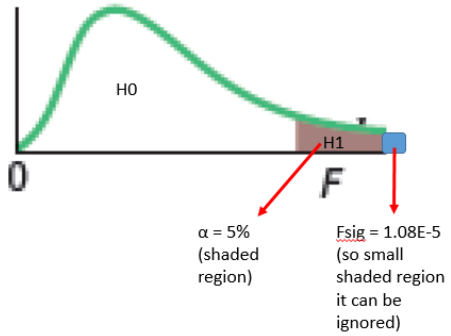
COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

I. SIGNIFICANCE F

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 20: Significance F

- Significance F (F_{sig}) is just another method for doing F Stat.
- Similarly (following after F Stat), if $\alpha = 5\%$
- & Significance F (F_{sig}) = 1.08E-05
- → $F_{sig} << \alpha$
- → Accept H1 → X is important.



COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

J. T STAT

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.959911184					
R Square	0.92142948					
Adjusted R Square	0.911608165					
Standard Error	26.50212256					
Observations	10					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	65895.2	65895.2	93.81936	1.07652E-05	
Residual	8	5618.9	702.3625			
Total	9	71514.1				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678

Figure 21: t stat

$$t = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}} = \frac{0.96\sqrt{10-2}}{\sqrt{1-(0.96)^2}} = \frac{2.7153}{0.28} = 9.690$$

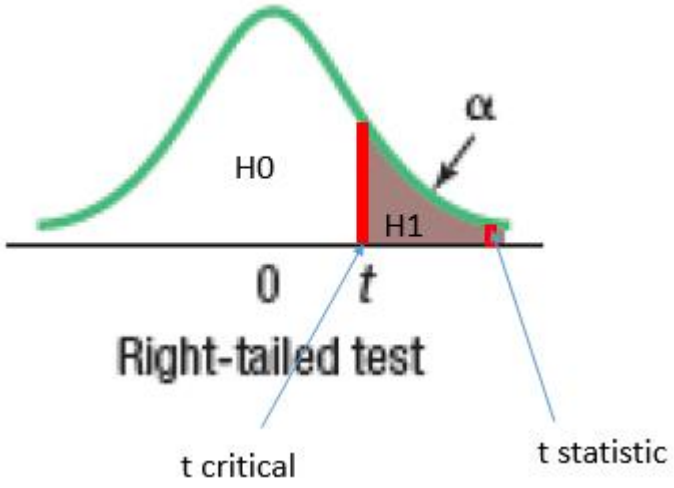
- t stat = 9.686
- t critical = 1.86
    - One-tailed test
    - $\alpha = 0.05$ significance level

- Refer to Table 5: Student's t distribution below.
- Since;
    - Null Hypothesis: $H_0 : \beta_1 = 0$ (X (age) is not important)

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

- Alternate Hypothesis: $H_1 : \beta_1 \neq 0$ (X(age) is important)

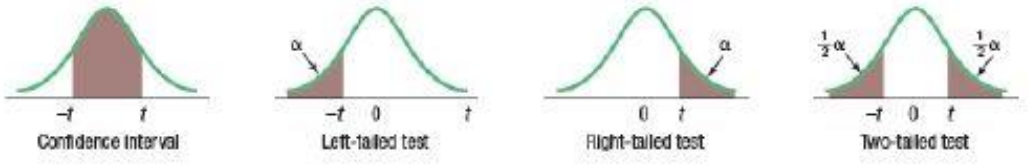
- t critical < t statistic → H1 is accepted.



COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

Table 5: Student's t distribution

B.2 Student's *t* Distribution



	Confidence Intervals, c					
	80%	90%	95%	98%	99%	99.9%
	Level of Significance for One-Tailed Test, α					
df	0.10	0.05	0.025	0.01	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
1	3.078	6.314	12.706	31.821	63.657	636.619
2	1.886	2.920	4.303	6.965	9.925	31.599
3	1.638	2.353	3.182	4.541	5.841	12.924
4	1.533	2.132	2.776	3.747	4.604	8.610
5	1.476	2.015	2.571	3.365	4.032	6.869
6	1.440	1.943	2.447	3.143	3.707	5.959
7	1.415	1.895	2.365	2.998	3.499	5.408
8	1.397	1.860	2.306	2.896	3.355	5.041
9	1.383	1.833	2.262	2.821	3.250	4.781
10	1.372	1.812	2.228	2.764	3.169	4.587
11	1.363	1.796	2.201	2.718	3.106	4.437
12	1.356	1.782	2.179	2.681	3.055	4.318
13	1.350	1.771	2.160	2.650	3.012	4.221
14	1.345	1.761	2.145	2.624	2.977	4.140
15	1.341	1.753	2.131	2.602	2.947	4.073
16	1.337	1.746	2.120	2.583	2.921	4.015
17	1.333	1.740	2.110	2.567	2.898	3.965
18	1.330	1.734	2.101	2.552	2.878	3.922
19	1.328	1.729	2.093	2.539	2.861	3.883
20	1.325	1.725	2.086	2.528	2.845	3.850
21	1.323	1.721	2.080	2.518	2.831	3.819
22	1.321	1.717	2.074	2.508	2.819	3.792
23	1.319	1.714	2.069	2.500	2.807	3.768
24	1.318	1.711	2.064	2.492	2.797	3.745
25	1.316	1.708	2.060	2.485	2.787	3.725
26	1.315	1.706	2.056	2.479	2.779	3.707
27	1.314	1.703	2.052	2.473	2.771	3.690
28	1.313	1.701	2.048	2.467	2.763	3.674
29	1.311	1.699	2.045	2.462	2.756	3.659
30	1.310	1.697	2.042	2.457	2.750	3.646
31	1.309	1.696	2.040	2.453	2.744	3.633
32	1.309	1.694	2.037	2.449	2.738	3.622
33	1.308	1.692	2.035	2.445	2.733	3.611
34	1.307	1.691	2.032	2.441	2.728	3.601
35	1.306	1.690	2.030	2.438	2.724	3.591

	Confidence Intervals, c					
	80%	90%	95%	98%	99%	99.9%
	Level of Significance for One-Tailed Test, α					
df	0.10	0.05	0.025	0.01	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
36	1.306	1.688	2.028	2.434	2.719	3.582
37	1.305	1.687	2.026	2.431	2.715	3.574
38	1.304	1.686	2.024	2.429	2.712	3.566
39	1.304	1.685	2.023	2.426	2.708	3.558
40	1.303	1.684	2.021	2.423	2.704	3.551
41	1.303	1.683	2.020	2.421	2.701	3.544
42	1.302	1.682	2.018	2.418	2.698	3.538
43	1.302	1.681	2.017	2.416	2.695	3.532
44	1.301	1.680	2.015	2.414	2.692	3.526
45	1.301	1.679	2.014	2.412	2.690	3.520
46	1.300	1.679	2.013	2.410	2.687	3.515
47	1.300	1.678	2.012	2.408	2.685	3.510
48	1.299	1.677	2.011	2.407	2.682	3.505
49	1.299	1.677	2.010	2.405	2.680	3.500
50	1.299	1.676	2.009	2.403	2.678	3.496
51	1.298	1.675	2.008	2.402	2.676	3.492
52	1.298	1.675	2.007	2.400	2.674	3.488
53	1.298	1.674	2.006	2.399	2.672	3.484
54	1.297	1.674	2.005	2.397	2.670	3.480
55	1.297	1.673	2.004	2.396	2.668	3.476
56	1.297	1.673	2.003	2.395	2.667	3.473
57	1.297	1.672	2.002	2.394	2.665	3.470
58	1.296	1.672	2.002	2.392	2.663	3.466
59	1.296	1.671	2.001	2.391	2.662	3.463
60	1.296	1.671	2.000	2.390	2.660	3.460
61	1.296	1.670	2.000	2.389	2.659	3.457
62	1.295	1.670	1.999	2.388	2.657	3.454
63	1.295	1.669	1.998	2.387	2.656	3.452
64	1.295	1.669	1.998	2.386	2.655	3.449
65	1.295	1.669	1.997	2.385	2.654	3.447
66	1.295	1.668	1.997	2.384	2.652	3.444
67	1.294	1.668	1.996	2.383	2.651	3.442
68	1.294	1.668	1.995	2.382	2.650	3.439
69	1.294	1.667	1.995	2.382	2.649	3.437
70	1.294	1.667	1.994	2.381	2.648	3.435

(continued)

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

	Confidence Intervals, c					
	80%	90%	95%	98%	99%	99.9%
	Level of Significance for One-Tailed Test, α					
df	0.10	0.05	0.025	0.01	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
71	1.294	1.667	1.994	2.380	2.647	3.433
72	1.293	1.666	1.993	2.379	2.646	3.431
73	1.293	1.666	1.993	2.379	2.645	3.429
74	1.293	1.666	1.993	2.378	2.644	3.427
75	1.293	1.665	1.992	2.377	2.643	3.425
76	1.293	1.665	1.992	2.376	2.642	3.423
77	1.293	1.665	1.991	2.376	2.641	3.421
78	1.292	1.665	1.991	2.375	2.640	3.420
79	1.292	1.664	1.990	2.374	2.640	3.418
80	1.292	1.664	1.990	2.374	2.639	3.416
81	1.292	1.664	1.990	2.373	2.638	3.415
82	1.292	1.664	1.989	2.373	2.637	3.413
83	1.292	1.663	1.989	2.372	2.636	3.412
84	1.292	1.663	1.989	2.372	2.636	3.410
85	1.292	1.663	1.988	2.371	2.635	3.409
86	1.291	1.663	1.988	2.370	2.634	3.407
87	1.291	1.663	1.988	2.370	2.634	3.406
88	1.291	1.662	1.987	2.369	2.633	3.405

	Confidence Intervals, c					
	80%	90%	95%	98%	99%	99.9%
	Level of Significance for One-Tailed Test, α					
df	0.10	0.05	0.025	0.01	0.005	0.0005
	Level of Significance for Two-Tailed Test, α					
	0.20	0.10	0.05	0.02	0.01	0.001
89	1.291	1.662	1.987	2.369	2.632	3.403
90	1.291	1.662	1.987	2.368	2.632	3.402
91	1.291	1.662	1.986	2.368	2.631	3.401
92	1.291	1.662	1.986	2.368	2.630	3.399
93	1.291	1.661	1.986	2.367	2.630	3.398
94	1.291	1.661	1.986	2.367	2.629	3.397
95	1.291	1.661	1.985	2.366	2.629	3.396
96	1.290	1.661	1.985	2.366	2.628	3.395
97	1.290	1.661	1.985	2.365	2.627	3.394
98	1.290	1.661	1.984	2.365	2.627	3.393
99	1.290	1.660	1.984	2.365	2.626	3.392
100	1.290	1.660	1.984	2.364	2.626	3.390
120	1.289	1.658	1.980	2.358	2.617	3.373
140	1.288	1.656	1.977	2.353	2.611	3.361
160	1.287	1.654	1.975	2.350	2.607	3.352
180	1.286	1.653	1.973	2.347	2.603	3.345
200	1.286	1.653	1.972	2.345	2.601	3.340
∞	1.282	1.645	1.960	2.326	2.576	3.291

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

K. P-VALUE

SUMMARY OUTPUT								
Regression Statistics								
Multiple R	0.959911184							
R Square	0.92142948							
Adjusted R Square	0.911608165							
Standard Error	26.50212256							
Observations	10							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	1	65895.2	65895.2	93.81936	1.07652E-05			
Residual	8	5618.9	702.3625					
Total	9	71514.1						
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678	43.73449322	71.06550678

Figure 22: P Value

- Notice P-Value for X (Age) = 1.08E-05 is exactly the same as Significance F (F_{sig}) = 1.08E-05
- This is because there is only one X (age).
- If there are multiple X's, they will be different.
- Since P-Value for X (Age) = 1.08E-05 << α = 5%
- Likewise accept H1 → X is important.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

L. LOWER AND UPPER 95%

SUMMARY OUTPUT						
Regression Statistics						
Multiple R	0.959911184					
R Square	0.92142948					
Adjusted R Square	0.911608165					
Standard Error	26.50212256					
Observations	10					
ANOVA						
	df	*SS*	*MS*	*F*	*Significance F*	
Regression	1	65895.2	65895.2	93.81936	1.07652E-05	
Residual	8	5618.9	702.3625			
Total	9	71514.1				
	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*
Intercept	77.5	19.65450012	3.943117328	0.004277	32.17664145	122.8233586
age (years)	57.4	5.926054758	9.686039421	1.08E-05	43.73449322	71.06550678

Figure 23: Lower and Upper 95%

- How did we get Lower 95% = 43.73 and Upper 95% = 71.066 for Age?
- The Mean Coefficient is 57.4.
- The 95% CI is = 57.4 – 43.73 = 13.67.
- Thus, the 95% CI is 57.4 ± 13.67 → (43.73, 71.07).

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

V. ASSUMPTIONS OF LINEAR REGRESSION

1. X values are independent.
2. Y is dependent on X.
3. Y values are Normally Distributed
4. Means of Y values lie on the Regression Line.
5. $S_{y.x}$ is the Std. Dev. of these Y values

   $\hat{Y} \pm 1 s_{y.x}$ encompasses about 68% of the observed values.

   $\hat{Y} \pm 2 s_{y.x}$ encompasses about 95% of the observed values.

   $\hat{Y} \pm 3 s_{y.x}$ encompasses virtually all of the observed values.

6. $S_{y.x}$ is a fixed constant.
7. There's no relationship between each Y value.
8. Each X has been picked independent of another X.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

- Linear Regression by hand is also known as the "Method of Least Squares".

Year	1	2	3	4	5	6	7
Number Sold	35	50	75	90	105	110	130

Figure 24: Example for Method of Least Squares

- Presume we want to obtain the Linear Equation for Figure 24.

$$T = b_0 + b_1 t$$

Where:

$$b_0 = \overline{Y} - b_1 \overline{t}$$

$$b_1 = \frac{\sum tY - \frac{\sum t \sum Y}{n}}{\sum t^2 - \frac{\left(\sum t\right)^2}{n}}$$

Figure 25: Formula for Method of Least Squares

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

Table 6: Applying the Method of Least Squares

	Year (t)	Numbers Sold (Y)	t2	tY
	1	35	1	35
	2	50	4	100
	3	75	9	225
	4	90	16	360
	5	105	25	525
	6	110	36	660
	7	130	49	910
Total:	28	595	140	2815
Average:	4	85		

$$b_1 = \frac{\sum tY - \frac{\sum t \sum Y}{n}}{\sum t^2 - \frac{\left(\sum t\right)^2}{n}}$$

$$= \frac{2815 - \frac{(28)(595)}{7}}{140 - \frac{28^2}{7}}$$

$$= 15.5357$$

$$b_0 = \overline{Y} - b_1\overline{t}$$

$$= 85 - 15.5357\,(4)$$

$$= 22.857$$

Hence Trend Line Equation:

$$T = b_0 + b_1 t$$

$$T = 22.857 + 15.536t$$

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG

Dr. Alvin Ang earned his Ph.D., Masters and Bachelor degrees from NTU, Singapore. He used to be a Scientist and Professor. He is currently an Entrepreneur and Business Consultant. More about him at his website www.AlvinAng.sg.

COPYRIGHTED BY DR ALVIN ANG
WWW.ALVINANG.SG